

Proportional Selection in Networks

Georgios Papasotiropoulos, Oskar Skibski, Piotr Skowron, and Tomasz Wąs

Abstract

We address the problem of selecting k representative nodes from a network, aiming to achieve two objectives: identifying the most influential nodes and ensuring the selection proportionally reflects the network’s diversity. We propose two approaches to accomplish this, analyze them theoretically, and demonstrate their effectiveness through a series of experiments.

1 Introduction

Consider the problem of selecting a fixed number of k nodes from a network. Our goal is twofold: to identify the most influential nodes, and to ensure that the selection proportionally represents the diversity within the network. For instance, consider a network composed of three groups of densely connected nodes. Assume the groups contain 50%, 30%, and 20% of all nodes, respectively, and connections between groups are relatively sparse. If the objective is to select $k = 10$ nodes, a proportional approach would involve selecting five most influential nodes from the first group, three from the second, and two from the third group. In this paper we design selection methods that capture this intuition, yet apply to networks with more complex structures.

Our model has broad applicability across various real-world scenarios. For instance, imagine a network where nodes represent political blogs or news websites, and the links indicate references between them. In this context, proportional selection would ensure a balanced representation of predominant opinions, such as left-wing and right-wing political views, within the chosen set of websites. As another example consider the case of a strike group described by Michael [29]: After acquiring a forest products facility, new management revised workers’ compensation, prompting a strike. When negotiations stalled, external consultants analyzed the workers’ communication network, revealing three distinct groups: young Spanish speakers, young English speakers, and older English speakers. Consultants identified Bob and Norm as key communicators and, by engaging them, quickly resolved the strike. Interestingly, for $k = 2$, our methods identify the same individuals. When $k = 3$, our best-performing method also selects Alejandro, ensuring representation from each of the predominant groups.

Finally, our model generalizes elections. As discussed in [Section 4](#), for bipartite graphs, it reduces to a well-studied election framework where proportionality is extensively explored [25, 17]. From what follows our work contributes to the social choice literature by enabling elections where the set of candidates is not predetermined. Instead, citizens can cast their votes for individuals they know personally. This concept has been previously explored in the context of vote delegation, or liquid democracy [21, 10]. However, liquid democracy often faces criticism because delegated votes tend to concentrate in the hands of a single influential individual. In contrast, our approach focuses on selecting a group of representatives based on votes cast through personal connections and trust. Thus, our approach combines the benefits of indirect democracy with a more personalized voting process, fostering a stronger connection between voters and their representatives.

Numerous measures of node importance has been proposed, and they are commonly known as centralities. In this paper, we propose two approaches that define selection rules for representative nodes based on given centrality measures. While our approaches are general and compatible with most prominent centrality measures as well as machine learning models, we focus here on two particular examples: PageRank [30] and Katz centrality [24]. Our first approach constructs elections from the underlying

graphs and applies contemporary methods from social choice theory. In the second approach, we evaluate subsets of nodes by modifying the original graph; specifically we prevent selected nodes from propagating their accumulated influence, hence limiting the extent to which a well-represented part of the network influences the assesment of remaining nodes. We analyze our methods axiomatically and characterize their behavior on special graph classes, where proportional selection is intuitively demonstrated. Our extensive experiments prove the potential for applying our methods in real scenarios.

Our work relates to research on group centrality measures [16, 3] and the strategic aspects of nodes selection in networks [2, 22, 5], yet, to the best of our knowledge, the notion of proportional representation has not been explored in either context. It also aligns with studies on proportional clustering [7, 13, 28], which focus on distance-based division of the nodes into groups rather than on selecting nodes based on their influence.

2 Preliminaries

We consider unweighted simple directed graphs. We will interpret nodes as voters and outgoing edges as their support for other nodes, or simply their votes. A *graph* (or a *network*) is a pair, $G = (V, E)$, where V is a set of n nodes and $E \subseteq V \times V$ is a set of m edges; the edges are ordered pairs of nodes. An edge (u, v) is an *outgoing edge* for node u and an *incoming edge* for v . The set of outgoing edges of u is denoted by $E^+(u)$. For a set of nodes $S \subseteq V$, we let $E^+(S) = \bigcup_{u \in S} E^+(u)$. The number of outgoing (resp. incoming) edges of u is called *out-degree* (resp. *in-degree*) of u and denoted by $\deg^+(u)$ (resp. $\deg^-(u)$).

A *walk* is a sequence of nodes $(v_1, \dots, v_k)$ such that every two consecutive nodes are connected by an edge: $(v_i, v_{i+1}) \in E$ for every $i \in \{1, \dots, k-1\}$. The length of such a walk is equal to $k-1$, i.e., the number of edges in the walk. Note that a single node is a walk of length 0. If in a walk all nodes are distinct, we call it a *path*. A node u is called a *predecessor* of w if there exists a walk $(v_1, \dots, v_k)$ such that $u = v_1$ and $w = v_k$. A node u is a *successor* of w if w is a predecessor of u . The set of predecessors of node u is denoted by $Pred(u)$ and its set of successors is denoted by $Succ(u)$. For a set of nodes $S \subseteq V$ we define $Succ(S) = \bigcup_{u \in S} Succ(u)$. The set of all walks in G is denoted by $\Omega(G)$.

For a set of nodes $S \subseteq V$, we denote by $G[S]$ the graph $(S, \{(u, v) \in E : u, v \in S\})$, i.e., the subgraph of G containing S and the edges between nodes from S . For a set of edges $M \subseteq E$, we denote by $G - M$ the graph $(V, E \setminus M)$, i.e., the subgraph of G that remains after the deletion of edges in M . A *component* of G (also known as *weakly connected component*) is a maximal subset of nodes $S \subseteq V$ such that every two nodes are connected by a walk in the *underlying* undirected graph inferred by G . A (sub)graph G is *strongly connected* if for every two nodes u, v of G there exists a walk from u to v . A *clique* is a graph in which every two nodes u, v are connected by an edge.

A graph $G = (V, E)$ is called (directed) *bipartite* if all walks in $\Omega(G)$ are of length of at most 1, in other words, if its set of nodes V can be divided into two disjoint sets $V = V_1 \cup V_2$ such that every edge in E is an outgoing for a node in V_1 and an incoming for a node in V_2 . A graph $G = (V, E)$ is called *functional* if $\deg^+(u) \leq 1$, for all nodes $u \in V$. The nomenclature arises from the observation that the set of edges can be interpreted as a function that assigns to each node (at most) one other.

Given a network $G(V, E)$ and an integer $k < n$ our goal is to select k nodes from V . A method that performs this selection is referred to as a *group selection rule for networks*, in short, a *rule*, and denoted by $\mathcal{R}$. We will also refer to the outcome of such a rule simply as $\mathcal{R}(G, k)$. In our pursuit to find a rule that elects most influential nodes in a proportional manner, we will use the measures of nodes' importance from network science as well as the concepts from social choice theory.

2.1 Centrality Measures

A *centrality measure* F is a function that for each graph $G = (V, E)$ and node $v \in V$ assigns a real value, denoted by $F_G(v)$. The higher the value, the more central the node is considered. Our methods can be combined with any centrality measure; yet, we will focus on centrality measures based on walks. At what follows, for notational simplicity, when it is clear from the context, we omit G and α from scripts.

PageRank [30]: For a given decay factor $\alpha \in (0, 1)$, PageRank of a node v in graph G is defined as:

$$\text{PR}_G^\alpha(v) = \sum_{(u_1, \dots, u_k, v) \in \Omega(G)} \frac{\alpha^k}{\prod_{i=1}^k \deg^+(u_i)}. \quad (1)$$

At a high level, PageRank of v is proportional to the expected number of times that v is visited by a random walk that starts from a random node and in each step follows an outgoing edge chosen uniformly at random or ends the walk with probability $1 - \alpha$. Hence, PageRank of a node mostly depends on the number and the importance of its predecessors. We note that many variants of PageRank appear in the literature (see, e.g., the work of Wąs and Skibski [39]).

Katz centrality [24]: For a given decay factor $\alpha \in (0, 1/\lambda)$, where λ is the largest eigenvalue of the adjacency matrix of G , Katz centrality of a node v in G is defined as:

$$\text{K}_G^\alpha(v) = \sum_{(u_1, \dots, u_k, v) \in \Omega(G)} \alpha^k. \quad (2)$$

Katz centrality can be seen as a parallel version of PageRank, where instead of counting the visits in a random walk we take the number of all walks that pass through a node (discounted by their length).

2.2 Election Rules

An (committee) election profile is a triple (V, C, μ) , where μ is a function that denotes the preferences expressed by the voters of V over the candidates of C . A voting rule is a function that for each profile (V, C, μ) and natural k , elects a committee consisting of k candidates of C according to the preferences of V as expressed by μ . Preferences can be expressed in various ways, e.g., using approval ballots where $\mu_i(c) = 1$ if voter i approves candidate c and $\mu_i(c) = 0$, otherwise, or using general utility functions in which $\mu_i(c)$ is an arbitrary non-negative value indicating the satisfaction of i from electing c . We particularly focus on scenarios where $V \equiv C$, i.e., voters aim to select a committee from among themselves by expressing preferences over one another.

Two of the simplest election rules are Approval Voting (AV) and Satisfaction Approval Voting (SAV). Under these rules, each candidate c is getting a score from each voter v_i , which equals $\mu_i(c)$ for AV and $\mu_i(c) / \sum_{c' \in C} \mu_i(c')$ for SAV. In both rules, the k candidates having the maximum total score form the winning committee. An alternative voting rule that aims to select sets of candidates in a proportional way has been recently proposed under the name Method of Equal Shares [33].

Method of Equal Shares (MES): Let b_i be a (virtual) budget of voter i , initially set to k/n . The rule operates in rounds. We say that a not yet elected candidate c is *ρ -affordable* for $\rho \in \mathbb{R}_+$, if its supporters can cover its (unit) cost in such a way that each of them pays ρ per unit of utility or all of their remaining funds. Hence, we calculate the minimum value of ρ that satisfies the expression $\sum_{i \in V} \min(b_i, \mu_i(c) \cdot \rho) \geq 1$. In a given round the method selects the candidate that is ρ -affordable for the lowest possible value of ρ and updates the voters' budgets accordingly: $b_i := b_i - \min(b_i, \mu_i(c) \cdot \rho)$ before proceeding to the next round. The procedure stops if there is no ρ -affordable candidate for any finite value of ρ .

This procedure may end with less than k candidates selected. To this end, in our experiments we will use the method with *AddIU completion method* [18]. Additionally, the recently proposed *Method of Equal Shares with Bounded Overspending (BOS)* [31], a robust variant of MES that balances proportionality and efficiency, will also be explored in our simulations.

3 Group Selection Rules for Networks

A natural approach to selecting influential nodes in a network is to choose the nodes of highest centrality. In the context of liquid democracy, Boldi et al. [8] proposed selecting the k nodes with the highest PageRank, a method we refer to as **TOPRANK**. This accounts for the diminishing trust along delegation paths. By **TOPKATZ**, we denote the analogous rule for Katz centrality. As we will show, these rules may severely fail to represent the network proportionally. In response, we propose two general approaches, which constitute our main conceptual contribution.

3.1 Election-Based Group Selection

The first approach leverages proportional committee election rules. At a high level, it transforms the input graph into an election scenario based on nodes' importance and applies a voting rule to select representatives. Specifically, for any two distinct nodes u and v , it defines a utility for node u derived from including node v in the selected set. Such an assessment can be derived from most centrality measures in a natural way, but can also be the result of a link prediction, similarity measures, or a machine learning model. For PageRank, with a decay factor α , we define $\mu_G^\alpha(u, v)$ as the expected number of visits at v of the random walk that starts at node u , based on Equation (1):

$$\mu_G^\alpha(u, v) := \sum_{(u_1, u_2, \dots, u_k, v) \in \Omega(G): u=u_1} \frac{\alpha^k}{\prod_{i=1}^k \deg^+(u_i)}. \quad (3)$$

According to this function, the utility is high if node v can be reached with high probability from u . If we interpret edges as votes' delegation, the node to which a vote can be delegated more directly is preferred. Note that $\text{PR}_G^\alpha(v) = \sum_{u \in V} \mu_G^\alpha(u, v)$, so **TOPRANK** is equivalent to AV rule applied to election (V, V, μ_G^α) . Instead, we use the Method of Equal Shares due to its well-known proportionality guarantees, resulting in a rule we refer to as **MESRANK**:

$$\text{MESRANK}^\alpha(G, k) = \text{MES}((V, V, \mu_G^\alpha), k).$$

For Katz centrality the definition of a utility function is analogous, but based on Equation (2), and this gives rise to a rule that we will call **MESKATZ**. PageRank and, in particular, the utility function from Equation (3), can be computed in polynomial time [26]. The same applies to the Method of Equal Shares. Hence, the outcome of **MESRANK** can be computed in polynomial time; the same holds for **MESKATZ**.

3.2 Absorbing Rules

The second approach we propose is directly inspired by a scenario where voters delegate their votes to others and we need to select a group, $S \subseteq V$, to represent the whole electorate. Clearly, the selected voters do not need to delegate their votes further (as they represent themselves), so we can say that they “absorb” the chain of delegations. Thus, for an arbitrary centrality measure F , where $F_G(i)$ indicates the centrality of node i in G , the importance of each of the selected nodes $i \in S$ can be defined as $F_{G-E^+(S)}(i)$, where the graph $G - E^+(S)$ is obtained by removing outgoing edges of all nodes from S . Now, we assess the group of selected nodes $S \subseteq V$ by the least important node: $F_G(S) = \min_{i \in S} F_{G-E^+(S)}(i)$. The idea is that every node in the selected group should have sufficiently large support from nodes not already represented by other nodes. The group selection rule chooses the group that maximizes this score. This approach combined with PageRank gives a rule, that we will call **ABSORBRANK**:

$$\text{ABSORBRANK}(G, k) = \arg \max_{S \subseteq V, |S|=k} \text{PR}_G^{\alpha \rightarrow 1}(S).$$

We maximize the decay factor by setting $\alpha \rightarrow 1$. In this way, we obtain an idealized version of PageRank that does not take the length of a walk into account.

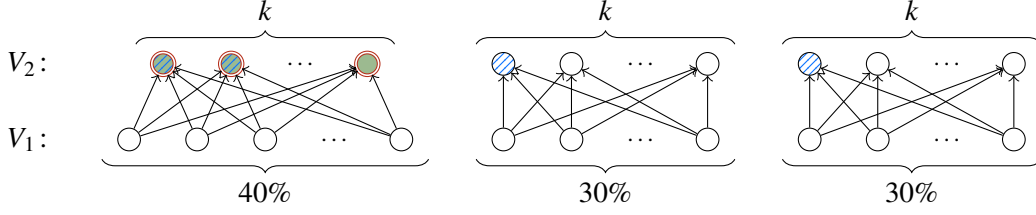


Figure 1: Selecting k nodes from a bipartite graph. TopRANK (red double lines) and AbsorbRank (green shading) select k nodes from the first group of V_2 . MesRANK (blue pattern) selects $0.4k$ of nodes from the first group and $0.3k$ from each other group of V_2 .

AbsorbKatz can be defined analogously, with the difference that we take $\alpha \rightarrow 1/\lambda$. We note that PageRank with $\alpha \rightarrow 1$ and Katz centrality with $\alpha \rightarrow 1/\lambda$ are closely related to other centrality measures called *Seeley index* [37] and *eigenvector centrality* [9], respectively; see [38] for a detailed discussion.

We prove that the outcomes of AbsorbRank and AbsorbKatz, unlike that of MesRank and MesKatz, cannot be computed in polynomial time (unless $P=NP$). This aligns with prior work identifying NP-hardness results for group centrality measures [40, 12, 4], making it of independent interest.

Theorem 1. Given an input (G, k) it is NP-hard to compute $\text{AbsorbRank}(G, k)$ and $\text{AbsorbKatz}(G, k)$.

We note that the proof of Theorem 1 does not rely on the fact that $\alpha \rightarrow 1$ and extends to all variants of AbsorbRank regardless of the choice of $\alpha \in (0, 1)$. In turn, given the intractability result, for our experiments, we will use heuristic approximations that take an arbitrary α and proceed sequentially. Specifically, $\text{SeqAbsorbRank}^\alpha$ works in k iterations, in each selecting node i that maximizes $\text{PR}_G^\alpha(S \cup \{i\})$, where S contains all nodes selected so far.

4 Case Studies

In this section, we illustrate our methods using two specific classes of graphs: bipartite and functional, as defined in Section 2. In these graph families proportionality can be intuitively captured, and hence they can be seen as a first way of examining proportionality achieved by the proposed methods and how they differentiate among them. Bipartite graphs mimic the scenario of representative democracy, where the set of candidates is separate from the set of voters. Functional graphs can be viewed as elections with 1-approval ballots (where each voter supports at most one candidate) and they serve as a widely studied case in the liquid democracy literature. They were also the exclusive focus of Boldi et al. [8]. Importantly, the outcome of all the examined rules can be computed in polynomial time in both of these graph families.

Theorem 2. $\text{AbsorbRank}(G, k)$ and $\text{AbsorbKatz}(G, k)$ can be computed in polynomial time when G is a functional or a bipartite graph.

4.1 Bipartite Graphs

We begin with directed bipartite graphs that are particularly able to highlight the differences between measures based on PageRank and Katz centrality, as well as the characteristic behavior of proportional election rules combined with network centrality measures. We assume that more than k nodes have incoming edges; otherwise, the examined problem becomes trivial. Since there are only walks of length 0 and 1, the Katz centrality and PageRank of each node v can be easily determined. Specifically, if v belongs to V_1 , its centrality is minimal, i.e., $\text{PR}_G^\alpha(v) = \text{K}_G^\alpha(v) = 1$. Conversely, if v belongs to V_2 , then:

$$\text{PR}_G^\alpha(v) = 1 + \sum_{(u,v) \in E} \frac{\alpha}{\text{deg}^+(u)}, \quad \text{K}_G^\alpha(v) = 1 + \alpha \cdot \text{deg}^-(v).$$

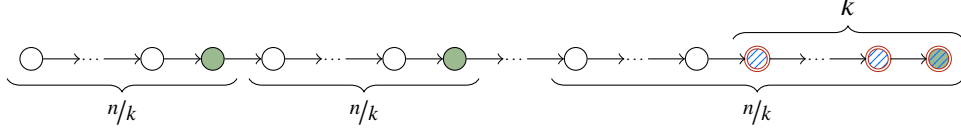


Figure 2: Selecting k nodes from the n -path (n divisible by k). TOP-RANK (red double lines) and MESRANK (blue pattern) choose the last k nodes. ABSORBRANK (green shading) selects nodes evenly splitting the path.

In words, nodes in V_2 get α per incoming edge under Katz centrality, and α divided by the number of outgoing edges of each supporting voter under PageRank. As a result, TOPKATZ simplifies to Approval Voting in this election instance, and TOPRANK simplifies to Satisfaction Approval Voting. We highlight that since "candidate" nodes in V_2 have no outgoing edges, ABSORBKATZ and ABSORBRANK yield the same results as TOPKATZ and TOPRANK, respectively.

The approach based on the Method of Equal Shares works differently. The utility that node u gets from selecting node v is $\mu_G^\alpha(u, v) = \alpha / \deg^+(u)$ for PageRank and $\mu_G^\alpha(u, v) = \alpha$ for Katz centrality. Consequently, MESKATZ corresponds to the result of running MES when all non-zero utilities are equal to α , while MESRANK splits α among all approved by u candidates. For a visualization, consider the bipartite graph depicted in Figure 1. Methods that select nodes based on the highest PageRank or Katz centrality will choose all k nodes from the first group of V_2 , ignoring the votes of the remaining 60% of voters. On the other hand, MESRANK and MESKATZ will select $0.4k$ nodes from the first group and $0.3k$ nodes from each other group of V_2 (up to rounding). This follows from the EJR property [33] saying that any large enough group approving the same set of candidates should get a proportional representation.

4.2 Functional Graphs

We now consider graphs in which every node has out-degree of at most one. PageRank and Katz centralities are identical on such graphs, hence, for ease of exposition, our analysis will center on rules based on PageRank. For simplicity, we assume that the decay factor approaches 1 for all considered rules: $\alpha \rightarrow 1$. In this case, if a node does not lie on any cycle, the PageRank of a node is nearly equal to its number of predecessors. More precisely, if $|Pred(u)| > |Pred(v)|$, then $PR_G^{\alpha \rightarrow 1}(u) > PR_G^{\alpha \rightarrow 1}(v)$.

We begin with an analysis of paths. Consider one of length n (say divisible by k). MESRANK views this instance as an election in which every node receives a utility of approximately 1 from each predecessor. Hence it selects the last k nodes of the path, as they have the maximum number of predecessors, clearly dominating the other nodes. The same holds for TOPRANK. However, ABSORBRANK behaves differently. The sink of the path will be again selected, but then, if the second-to-last node is selected, the sink will have no incoming edges in $G - E^+(S)$, and its PageRank will be minimized. Therefore, the rule avoids such a selection. Instead, it aims to select nodes in such a way that they have equal support in the graph without their outgoing edges: it selects nodes $n/k, 2 \cdot n/k, \dots, n$. For an illustration see Figure 2.

Consider an arbitrary (connected) functional graph. Such a graph consists of at most one cycle and in-trees attached to nodes from the cycle. Nodes from the cycle clearly have the maximal PageRank and have non-zero (close to 1) utility for all the nodes. Thus, if the cycle contains at least k nodes, then both TOPRANK and MESRANK would select only nodes from the cycle. If not, both methods select all nodes from the cycle and then some nodes from the attached in-trees. This is where both methods begin to differ; see Figure 3 for an example. TOPRANK will select nodes with the highest number of predecessors. All such nodes may come from the same in-tree. However, from each in-tree MESRANK will select a number of nodes proportional to its size. We switch our attention to ABSORBRANK. If the graph has a cycle, then ABSORBRANK would select one node from there, say v . The graph $G - E^+(v)$ is a tree. Hence, roughly speaking, ABSORBRANK splits this tree into parts of equal size, selecting the root from each.

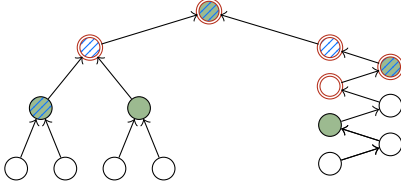


Figure 3: Selecting $k = 5$ nodes from a (directed in-)tree with two unbalanced branches of equal size. TOPRANK (red double lines) selects mostly from the right-hand side. MESRANK (blue pattern) selects two nodes from each side. ABSORBRANK (green shading) splits the tree in subtrees of size 3.

Method	Entitlement		
	Clique-	Component-	Subgraph-
TOPKATZ/TOPRANK	✗	✗	✗
ABSORBKATZ	✗	✗	✗
ABSORBRANK	✓	✗	✗
MESKATZ/MESRANK	✓	✓	✓

Table 1: Summary of the axiomatic properties.

5 Axioms of Proportionality

We now introduce axioms to formalize the intuition that a sufficiently large and cohesive group of nodes deserves a proportional number of representatives. Our results (depicted in Table 1) shed further light on the differences among the examined rules. Specifically, our axioms are inspired by the literature on multiwinner election rules. Following the approach from works in this area (e.g., [6]), the idea is that each node should have a significant influence over k/n of the committee. Consequently, a cohesive group S is entitled to $\lfloor k \cdot |S|/n \rfloor$ representatives. The key question, then, is: in the context of our study, which groups of nodes can be considered cohesive, and which nodes qualify as proper representatives of these groups?

We view a group of nodes as cohesive if all nodes mutually approve each other, either directly or indirectly. The most cohesive groups of nodes possible are the ones that form a component that is a clique. Our first axiom states that each such a component is entitled to a representation proportional to its size.

Clique-Entitlement: For every graph $G = (V, E)$, if there exists a component $S \subseteq V$ such that $G[S]$ is a clique, then $|S \cap \mathcal{R}(G, k)| \geq \lfloor k \cdot |S|/n \rfloor$.

TOPRANK and TOPKATZ do not satisfy this axiom, as they may overlook nodes from a clique component that should receive a representation when a larger and more diverse component exists. In contrast, we will later show that ABSORBRANK, MESRANK, and MESKATZ satisfy this axiom.

Proposition 3. TOPRANK and TOPKATZ do not satisfy Clique-Entitlement.

Next, we generalize Clique-Entitlement and require only that the component is strongly connected. It turns out that ABSORBRANK doesn't satisfy this stronger axiom, in contrast to MESRANK and MESKATZ.

Component-Entitlement: For every graph $G = (V, E)$, if there exists a component $S \subseteq V$ that is strongly connected, then $|S \cap \mathcal{R}(G, k)| \geq \lfloor k \cdot |S|/n \rfloor$.

Recall that ABSORBRANK attempts to divide the graph G into equal parts to maximize the minimal PageRank in $G - E(S)$ among nodes in S . Given this, ABSORBRANK violates Component-Entitlement because, depending on the structure, some components may be easier or harder to divide into several equal parts. As a result, it might be more beneficial to select an additional node from one component rather than another. Interestingly, ABSORBKATZ not only fails this axiom but also Clique-Entitlement, which highlights a crucial difference between the two absorbing methods.

Theorem 4. ABSORBRANK satisfies Clique-Entitlement, but does not satisfy Component-Entitlement. ABSORBKATZ does not satisfy Clique-Entitlement.

On the other hand, MESRANK and MESKATZ also satisfy a stronger property. Consider an arbitrary strongly connected subgraph within a larger component of a graph. Since this subgraph is not entirely separate

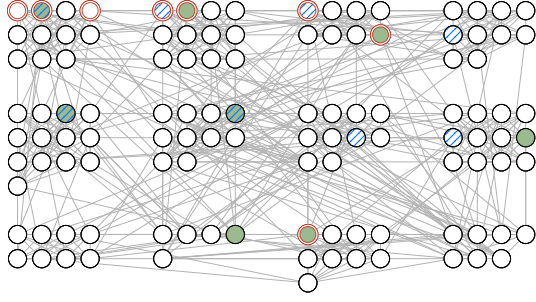


Figure 4: The College Football Network, where each group of nodes represents one of 11 conferences or a group of independent teams. For $k = 8$, TOPRANK (red double lines) selects nodes only from 4 conferences, while MESRANK (blue pattern) and SEQABSORBRANK (green shading) select at most 1 team per conference.

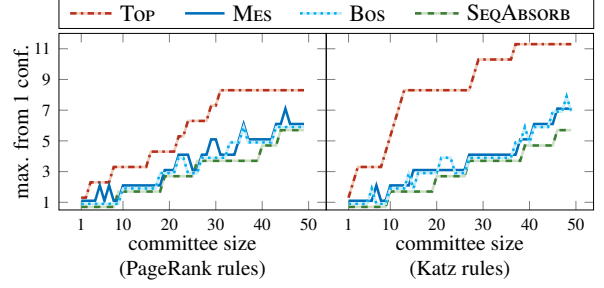


Figure 5: Maximum number of nodes from one conference that are selected by our rules for a given committee size in the College Football Network.

from other nodes, its delegated votes may flow outside the group. However, these votes will always flow to their successors. In the previous axioms, we assumed that a group deserving representation would be represented by its own members. In contrast, we now allow for representation through successors.

Subgraph-Entitlement: For every graph $G = (V, E)$, if there exists a subset of nodes $S \subseteq V$ such that $G[S]$ is strongly connected, then $|(S \cup \text{Succ}(S)) \cap \mathcal{R}(G, k)| \geq \lfloor k \cdot |S|/n \rfloor$.

Subgraph-Entitlement implies Component-Entitlement, and, in turn, Clique-Entitlement. Consequently, Subgraph-Entitlement is violated by TOPRANK, ABSORBRANK, and their variants. Our positive result for MESRANK and MESKATZ is even more general, as it applies to all selection rules based on the Method of Equal Shares and any centrality measure that satisfies a basic consistency condition.

Theorem 5. Assume that a utility function μ_G satisfies $\mu_G(u, v) > 0$ if and only if there is a path from u to v . Applying the Method of Equal Shares for such a utility function results in a rule satisfying Subgraph-Entitlement. In particular, MESRANK and MESKATZ satisfy this property.

6 Experiments on Real-Life and Synthetic Data

We compare TOPRANK, SEQABSORBRANK, and MESRANK, along with their Katz counterparts, empirically. Additionally, we include BosRANK and BosKATZ, defined similarly to MESRANK and MESKATZ but using the fine-tuned variant of MES called the Method of Equal Shares with Bounded Overspending (BOS) [31]. This is argued to better handle data with high variance in candidate utilities, a common characteristic of our model. We set $\alpha = 0.85$ for PageRank-based rules, as used by [11], and $\alpha = 0.85/\lambda$ for Katz-based ones. First, we analyze two network datasets often used as benchmarks for community detection [23].

6.1 College Football Network

The first dataset [20] is a graph of 115 nodes representing U.S. college football teams. Each undirected edge denotes a game played in Division IA during the 2000 Fall season (see Figure 4). We interpret each edge as a pair of directed edges in both directions. The teams are split into 11 conferences and a group of independents. Around 64% of games occur within the conferences and the rest are played across them. Figure 4 shows the outcomes of TOPRANK, MESRANK, and SEQABSORBRANK for $k = 8$. To generalize this analysis for other values of k and the Katz versions of our rules, we plot the maximum number of teams selected from a single conference for each rule and each $k \in \{1, \dots, 50\}$ (see Figure 5). Katz and PageRank yield similar results: TOP approaches select the most from one conference, while SEQABSORB select the fewest. MES and Bos closely align with SEQABSORB.

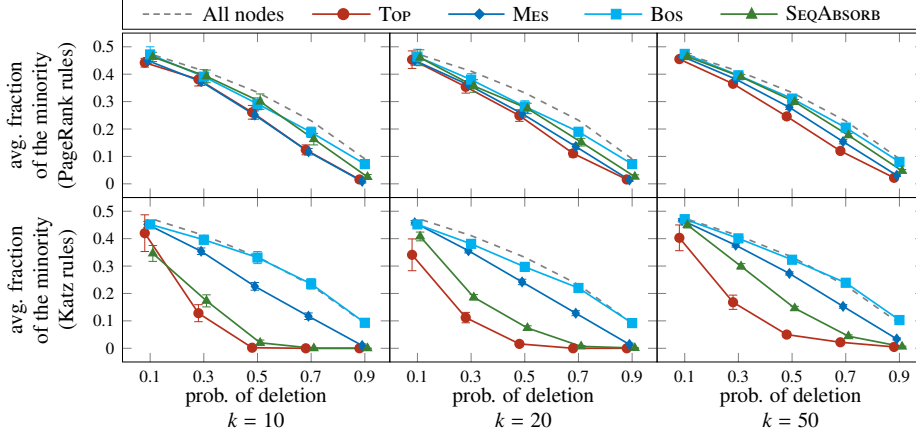


Figure 6: The fraction of nodes with the minority label in the graph (gray dashed lines) and in the rule outputs (lines with markers) in the Political Blogs Network after removing blogs from one side of the spectrum. Vertical lines denote 95% confidence intervals.

6.2 Political Blogs Network

The second network [1] consists of 1,490 nodes representing political blogs active during the 2004 U.S. presidential election. A directed edge from blog A to blog B indicates a front-page link from A to B. Blogs are labeled as “liberal” or “conservative” with a roughly balanced composition: 758 liberal and 732 conservative. To assess how our rules perform on unbalanced data, we delete each node of a given fixed label with probability p and evaluate how well the label distribution in the outcomes reflects that in the modified graph. For each $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, we generate 100 graphs (50 graphs with liberal nodes removed, and 50 with conservative). Figure 6 shows the fraction that the nodes from the pruned side constitute among all nodes (grey dashed line) and among nodes selected by our rules. The results are consistent across different values of k . For PageRank-based rules, the proportions of nodes with the two labels in the input graphs are closely reflected in the outcomes of our rules. We see this especially for lower deletion probabilities. At $p = 0.9$, BosRANK performs slightly better than the other rules. For Katz-based rules, the differences are stark. BosKATZ closely follows the proportions in the input graphs, outperforming BosRANK in this aspect. MESKATZ performs a bit worse, but TOPKATZ and ABSORBKATZ exclude minority nodes even for moderate values of p .

6.3 Further Real-World Data

We analyze two additional real-world networks, each with nodes split into multiple categories. The first one is a network of Facebook pages, where nodes represent official pages and edges denote mutual likes [36]. It includes pages from four categories: politicians, governmental organizations, TV shows, and companies. The second one is a CiteSeer network where nodes represent articles labeled with six academic categories [35]. For both datasets, in Table 2, we report the ℓ_1 distance between the vector of frequencies of different node labels in the outputs of our rules for $k \in \{10, 20, 50\}$ and the corresponding vector of frequencies for the entire network. Note that for the Facebook network and $k = 10$, the distance of BosRANK is larger than that of TOPRANK, MESRANK, and SEQABSORBRANK. Two factors might have resulted in that. First, for $k = 10$, selecting one node instead of another may change the distance significantly, hence there is a high degree of randomness in the outcomes. Second, 0.79 is not a particularly bad outcome, but the other rules, including TOPRANK, seem to perform quite well in this case. This is due to the fact that PageRank itself gives somewhat proportional outcomes; refer to Section 6.5 for details.

network	k	TOPRANK	MESRANK	BOSRANK	SEQABSRANK	TOPKATZ	MESKATZ	BOSKATZ	SEQABSKATZ
Facebook	10	0.68	0.41	0.79	0.67	1.39	0.99	0.59	1.39
	20	0.58	0.49	0.39	0.59	1.39	0.99	0.49	1.39
	50	0.51	0.42	0.27	0.51	1.35	0.59	0.39	1.15
CiteSeer	10	0.83	1.03	0.44	0.83	1.59	1.23	0.41	1.59
	20	0.53	0.73	0.42	0.43	1.59	0.83	0.28	1.49
	50	0.35	0.55	0.22	0.16	1.23	0.67	0.17	1.11

Table 2: The ℓ_1 distance between the vector of frequencies of different node labels in the outputs of our rules and the corresponding vector of frequencies for the entire network. The smallest distance among rules based on a particular centrality is highlighted.

6.4 Euclidean Data

Our final set of experiments is inspired by applications in social choice. Our methods can be applied to elect a committee among a group of voters who vote amongst themselves (without a predetermined set of candidates). In the social choice literature, voters and candidates are often modeled as points in a two-dimensional Euclidean space, typically representing an ideological spectrum [15]. Similarly, we assume that the nodes in our graph (representing both voters and candidates) correspond to points. These points are sampled from a specific distribution, which we describe later. Edges between nodes are introduced based on the distances between them, using one of four strategies. The first two strategies align with conventional assumptions in social choice, where voters are more likely to prefer candidates closer to them in the ideological space:

E-radius: Each node connects (with a certain fixed probability) to the nodes within a specified radius.

E-appr: Each node connects to a fixed number of its closest neighbors. To introduce some noise, we assume that each neighbor can be omitted with a fixed probability.

We also propose a novel Euclidean model, which is well aligned with the idea inspired by liquid democracy: that the voters tend to vote for their close friends (i.e., close points), yet in their preferences, exhibit a bias toward candidates with higher competence. In this model, each candidate is assigned an objective value representing their competence (represented as the y -coordinate). A node located at point (x, y) connects to the nodes closest to $(x, y + b)$ rather than to itself, where b is a constant representing the competence bias. Variants of this model are denoted as **B-radius** and **B-appr**, respectively, depending on whether the voters tend to approve the candidates within a certain radius of acceptability or a certain number of candidates.

For each of these models, the points are drawn from two Gaussian distributions, with the points divided between the two groups in a 1 : 3 or 2 : 3 ratio. From each setting, we sample 1000 instances. Each instance consists of $n = 1000$ points. For each instance, we construct a graph with n nodes and select $k = 10$ of them using one of our four methods combined with PageRank or Katz centrality. We identify the points corresponding to the selected nodes and mark them (in green) on the plot. The points from 1000 experiments are combined into a single plot, forming a histogram that represents the distribution of selected points. The histograms are presented in Figure 7 and Figure 8. The first row shows the distribution of the points, while the subsequent rows illustrate the histograms generated for our rules. Additionally, below each histogram, we provide the proportion of points selected from each half of the histogram. First, we observe that BosKATZ and BosRANK most accurately reflect the original proportions of the points in the two Gaussians; other PageRank rules and MesKATZ perform slightly worse, while ABSORBKATZ performs significantly worse, particularly when edges are formed based on radii. TopKATZ often fails to provide any proportionality, which strongly motivates our proposed methods. Second, we confirm that our methods recover competence bias from the graph, selecting nodes corresponding to high competence. Interestingly, these are not necessarily the nodes with the highest in-degree.

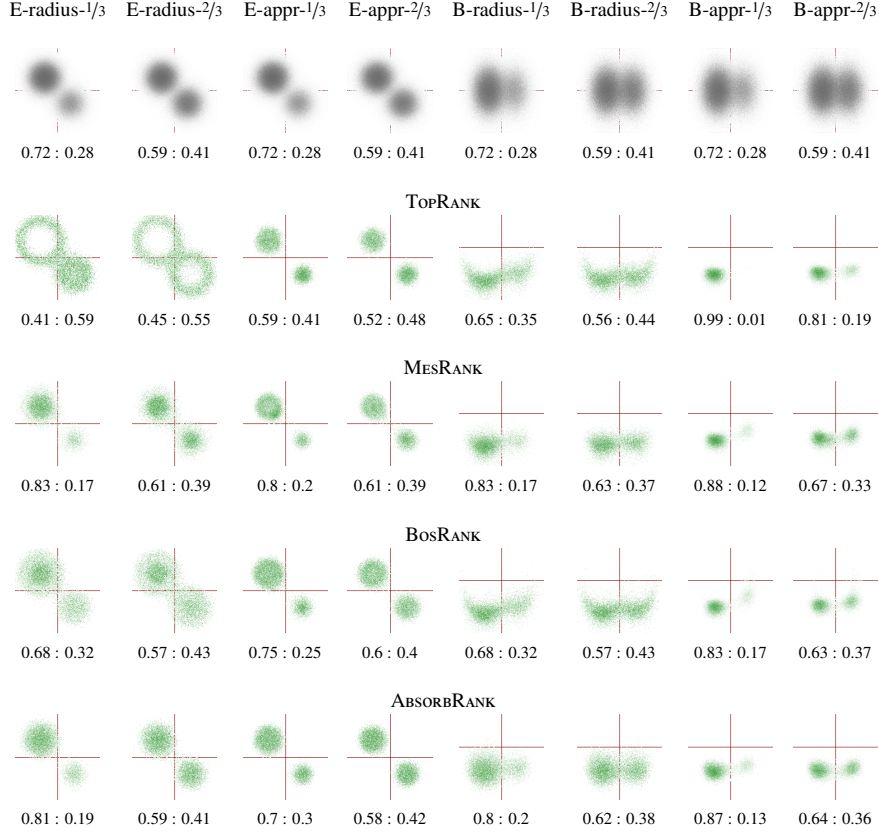


Figure 7: Histograms generated by our PageRank-based methods for Euclidean graphs. The distributions (first row) are presented at their original scale, while the histograms are displayed at 130% zoom for improved clarity.

6.5 Outlook of the Experiments

As a general rule, we observe that the straightforward approach of selecting the k nodes with the highest centralities leads to a more proportional representation under PageRank than under Katz centrality. This suggests that PageRank is inherently more proportional, which is also reflected in our axiomatic analysis (Section 5). To some extent, this can be explained by the fact that in PageRank every node distributes its contribution evenly among the centralities of other nodes, leaving no possibility of increasing one’s impact. In particular, PageRank significantly limits the influence of even the most influential nodes with higher out-degree. The strict dampening of influence based on out-degree is often undesirable, making PageRank unsuitable for certain scenarios. For example, if we interpret edges as votes, PageRank extends satisfaction approval voting, while Katz follows the logic of approval voting (see Section 4.1). In social choice theory, approval voting is often considered the preferable method and is widely used in practice [27]. This observation is further supported by the first two columns of Figure 7, where PageRank tends to select more extremist nodes in the surrounding of the two Gaussians. This occurs because these nodes receive edges from nodes with a very few outgoing edges on their own, which greatly boosts PageRank. In a way the candidates coming from less popular regions of voters preferences are additionally privileged, an arguably very undesired behaviour in election context. Additionally, the axiomatic analysis of PageRank per se suggests that it is unsuitable for certain applications [39]. In such applications, our work mitigates the limitations of PageRank by allowing the use of Katz.

We observe that all of our proposed methods enable a significantly more representative selection of nodes compared to the Top approach. The difference is particularly pronounced for Katz centrality, but even for PageRank we see noticeable improvements across different datasets. This effect is especially

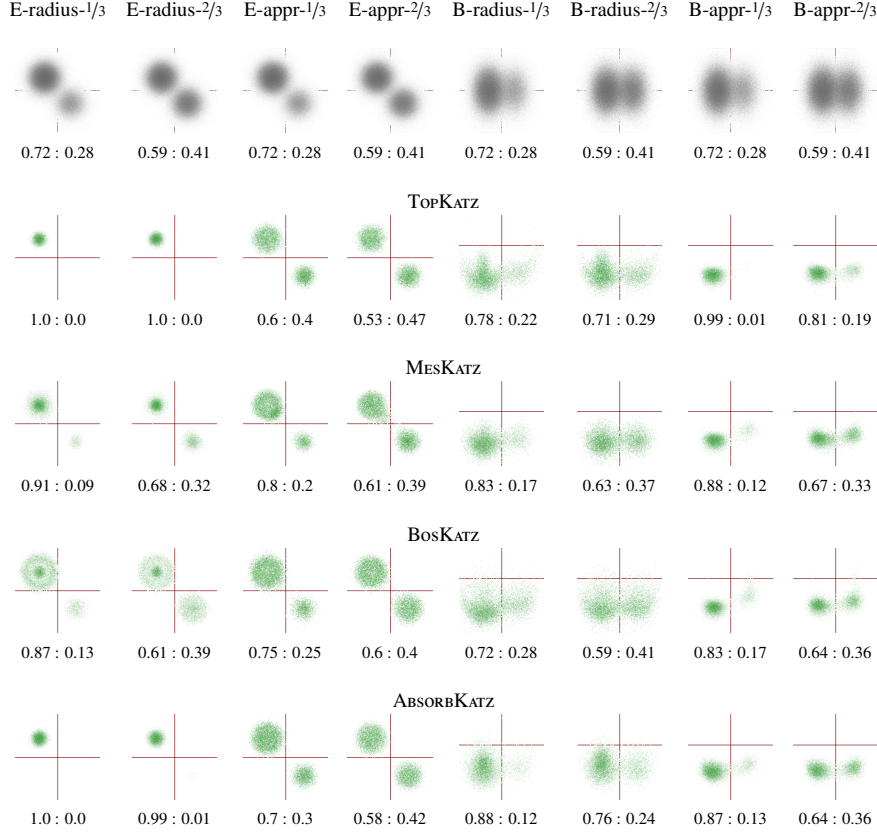


Figure 8: Histograms generated by our Katz-based methods for Euclidean graphs. The distributions (first row) are presented at their original scale, while the histograms are displayed at 150% zoom for improved clarity.

evident in Euclidean graphs (particularly in columns 1, 2, 7, and 8 of Figure 7) and in the College Football Network. Among the studied rules, those based on the Method of Equal Shares with Bounded Overspending—namely BosRANK and BosKATZ—yield the most representative committees, regardless of whether they are combined with PageRank or Katz centrality. Even when these methods do not yield the best overall results, they never perform poorly in terms of proportionality in any of our examined scenarios. The absorbing rules, on the other hand, prove more effective when used with PageRank rather than Katz, which is being also justified in our axiomatic analysis (Section 5). We highlight that this approach interprets proportionality differently, prioritizing representatives who minimize their maximum length of the influence propagation chain to non-selected voters, even if this means overlooking nodes that are highly influential in the original graph (see Section 4.2). The universality of the Bos approach suggests that it is also a preferable method to be combined with centrality measures beyond the ones studied in this paper as well as machine learning models.

7 Conclusion

We introduced the problem of selecting k influential nodes from a network while ensuring proportional representation of different groups implicitly present in the structure of the network. We proposed two techniques of extending a given centrality measure to a proportional node-selection method, focusing on the PageRank and Katz centralities. Our theoretical studies and experimental analyses yield largely similar conclusions. Both of our techniques show a significant improvement over the TOP approach. The approach based on the Method of Equal Shares (MES) offers the strongest axiomatic guarantees, and its specific variant incorporating the Method of Equal Shares with Bounded Overspending (BOS) performs best across the datasets used in our evaluation.

Acknowledgments. T. Waş was partially supported by EPSRC under grant EP/X038548/1. The rest of the authors were partially supported by the European Union (ERC, PRO-DEMOCRATIC, 101076570). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.



References

- [1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *Proceedings of the International Workshop on Link Discovery*, page 36–43, 2005.
- [2] N. Alon, F. Fischer, A. Procaccia, and M. Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110, 2011.
- [3] E. Angriman, A. van der Grinten, A. Bojchevski, D. Zügner, S. Günnemann, and H. Meyerhenke. Group centrality maximization for large-scale graphs. In *Proceedings of the workshop on Algorithm Engineering and Experiments*, pages 56–69, 2020.
- [4] E. Angriman, R. Becker, G. d’Angelo, H. Gilbert, A. van Der Grinten, and H. Meyerhenke. Group-harmonic and group-closeness maximization—approximation and engineering. In *Proceedings of the Workshop on Algorithm Engineering and Experiments*, pages 154–168, 2021.
- [5] H. Aziz, O. Lev, N. Mattei, J. Rosenschein, and T. Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 397–403, 2016.
- [6] H. Aziz, M. Brill, V. Conitzer, E. Elkind, R. Freeman, and T. Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485, 2017.
- [7] H. Aziz, B. E. Lee, S. M. Chu, and J. Vollen. Proportionally representative clustering. In *Proceedings of the Conference on Web and Internet Economics*, 2024.
- [8] P. Boldi, F. Bonchi, C. Castillo, and S. Vigna. Viscous democracy for social networks. *Communications of the ACM*, 54(6):129–137, 2011.
- [9] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113–120, 1972.
- [10] M. Brill. Interactive democracy. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems (Blue Sky Ideas Track)*, pages 1183–1187, 2018.
- [11] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.
- [12] C. Chen, W. Wang, and X. Wang. Efficient maximum closeness centrality group identification. In *Proceedings of the Australasian Database Conference*, pages 43–55, 2016.
- [13] X. Chen, B. Fain, L. Lyu, and K. Munagala. Proportionally fair clustering. In *Proceedings of the International Conference on Machine Learning*, pages 1032–1041, 2019.
- [14] S. Dasgupta, C. H. Papadimitriou, and U. Vazirani. *Algorithms*. McGraw-Hill, Inc., 2006.
- [15] E. Elkind, P. Faliszewski, J. Laslier, P. Skowron, A. Slinko, and N. Talmon. What do multiwinner voting rules do? An experiment over the two-dimensional Euclidean domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 494–501, 2017.
- [16] M. G. Everett and S. P. Borgatti. The centrality of groups and classes. *Journal of mathematical sociology*, 23(3):181–201, 1999.

- [17] P. Faliszewski, P. Skowron, A. Slinko, and N. Talmon. Multiwinner voting: A new challenge for social choice theory. In *Trends in Computational Social Choice*. AI Access, 2017.
- [18] P. Faliszewski, J. Flis, D. Peters, G. Pierczynski, P. Skowron, D. Stolicki, S. Szufa, and N. Talmon. Participatory budgeting: Data, tools and analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2667–2674, 2023.
- [19] G. N. Frederickson. Optimal algorithms for tree partitioning. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 168–177, 1991.
- [20] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [21] J. Green-Armytage. Direct voting and proxy voting. *Constitutional Political Economy*, 26:190–220, 2015.
- [22] R. Holzman and H. Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.
- [23] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, S. Y. Philip, and W. Zhang. A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1149–1170, 2021.
- [24] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
- [25] M. Lackner and P. Skowron. *Multi-Winner Voting with Approval Preferences*. Springer Briefs in Intelligent Systems. Springer, 2023.
- [26] A. N. Langville and C. D. Meyer. *Google’s PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2006.
- [27] J.-F. Laslier and M. R. Sanver, editors. *Handbook on Approval Voting*. Springer, 2010.
- [28] B. Li, L. Li, A. Sun, C. Wang, and Y. Wang. Approximate group fairness for clustering. In *Proceedings of the International Conference on Machine Learning*, pages 6381–6391, 2021.
- [29] J. H. Michael. Labor dispute reconciliation in a forest products manufacturing facility. *Forest Products Journal*, 47(11/12):41, 1997.
- [30] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.
- [31] G. Papasotiropoulos, S. Z. Pishbin, O. Skibski, P. Skowron, and T. Wąs. Method of equal shares with bounded overspending. *arXiv preprint arXiv:2409.15005*, 2024.
- [32] Y. Perl and S. R. Schach. Max-min tree partitioning. *Journal of the ACM*, 28(1):5–15, 1981.
- [33] D. Peters and P. Skowron. Proportionality and the limits of welfarism. In *Proceedings of the ACM Conference on Economics and Computation*, pages 793–794, 2020.
- [34] D. Peters, G. Pierczyński, and P. Skowron. Proportional participatory budgeting with additive utilities. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 12726–12737, 2021.
- [35] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4292–4293, 2015. <https://networkrepository.com>.
- [36] B. Rozemberczki, C. Allen, and R. Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2), 2021.
- [37] J. R. Seeley. The net of reciprocal influence. a problem in treating sociometric data. *Canadian Journal of Experimental Psychology*, 3:234, 1949.
- [38] T. Wąs and O. Skibski. An axiom system for feedback centralities. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 443–449, 2021.
- [39] T. Wąs and O. Skibski. Axiomatic characterization of pagerank. *Artificial Intelligence*, 2023. 103900.
- [40] J. Zhao, J. C. Lui, D. Towsley, and X. Guan. Measuring and maximizing group closeness centrality over disk-resident graphs. In *Proceedings of the International Conference on World Wide Web*, pages 689–694, 2014.

Georgios Papasotiropoulos
University of Warsaw
Warsaw, Poland
Email: gpapasotiropoulos@gmail.com

Oskar Skibski
University of Warsaw
Warsaw, Poland
Email: oskar.skibski@mimuw.edu.pl

Piotr Skowron
University of Warsaw
Warsaw, Poland
Email: p.skowron@mimuw.edu.pl

Tomasz Wąs
University of Oxford
Oxford, United Kingdom
Email: tomasz.was@cs.ox.ac.uk

Appendix

A Omitted Proofs

In this appendix, we present the proofs of our theorems and propositions.

A.1 Proof of Theorem 1

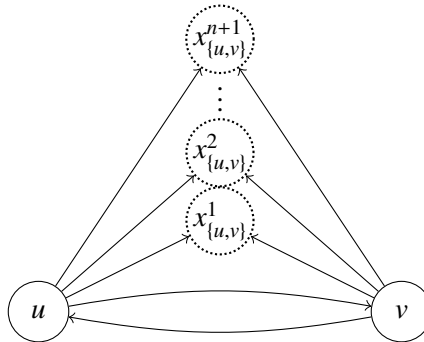
First we define the decision version of the computational problem that corresponds to **ABSORBRANK**, which we are going to prove hardness for. The problem statement and the proof for **ABSORBKATZ** are analogous.

ABSORBRANK	
Input:	A directed graph $G = (V, E)$ and parameters $c \in \mathbb{R}_{\geq 0}, k \in \mathbb{N}$.
Question:	Does there exist a set $S \subseteq V$ such that $ S = k$ and $\text{PR}_{G-E^+(S)}^{\alpha \rightarrow 1}(S) \geq c$?

We reduce from the Independent Set (IS) problem, where, given an undirected graph $G' = (V', E')$ and an integer r , the goal is to determine if there exists an *independent set* of size r , i.e., a subset of nodes $R \subseteq V'$ such that $|R| = r$ and no two nodes in R are connected by an edge in G' . The NP-hardness holds even when G' has no nodes of degree 0 or 1 [14]. Given an instance of IS, $I' = (G', r)$, with n nodes in V' and m edges in E' , we create an instance $I = (G, c, k)$ of **ABSORBRANK** with graph $G = (V, E)$ as follows:

- We let the set of nodes V contain all *original* nodes from V' as well as $n + 1$ *new* nodes for each edge in E' , i.e., $V = V' \cup V_{\text{new}}$, where $V_{\text{new}} = \{x_e^i : e \in E', i \in [n + 1]\}$. This gives us a set V of $m(n + 1) + n$ nodes in total.
- For each undirected edge $\{u, v\} \in E'$, let E contain a pair of directed edges (u, v) and (v, u) as well as an edge from each of the original nodes u and v to each new node corresponding to edge $\{u, v\}$. Formally, we have $E = \bigcup_{\{u,v\} \in E'} (\{(u, v), (v, u)\} \cup \{(u, x_{\{u,v\}}^i), (v, x_{\{u,v\}}^i) : i \in [n + 1]\})$.
- Finally, we set $k = r + m(n + 1)$ and $c = 1 + \epsilon$, where ϵ is an arbitrary constant smaller than $1/2 \cdot \max_{u \in V} \deg_G^+(u)$.

The following figure depicts the used gadget for an edge (u, v) of G' . The nodes with the dotted lines correspond to the nodes termed as new, while the remaining two nodes correspond to the original ones.



Observe that if $\alpha \geq 1/2$, a node v can have the value of PageRank less than c in a subgraph G'' of G only if it does not have any incoming edges in G'' . Indeed, if it has at least one incoming edge, say from u , then $\text{PR}_{G''}^\alpha(v) \geq 1 + \alpha / \deg_{G''}^+(u) \geq c$. Thus, the question of the decision variant of **ABSORBRANK** in the instance I can be equivalently expressed as follows: *Does there exist a subset $S \subseteq V$ with $|S| = k$ such that $\deg_{G-E^+(S)}^-(i) > 0$ for every $i \in S$ (i.e., there is no node in S that has a zero in-degree after the*

removal of the outgoing edges of nodes in S)? Let us show that the positive answer to this question for I is equivalent to the existence of an independent set of size r in I' .

First, assume that there is an independence set R of size r in G' . We will show that this implies that $S = R \cup V_{new}$, i.e., the subset of the original nodes corresponding to R together with all the new nodes, witnesses that I is a yes-instance as well. Observe first that $|S| = r + m(n+1) = k$. Thus, it remains to show that every node in S has a positive in-degree in the graph $G - E^+(S)$.

Fix an arbitrary $i \in S$. If $i \in V_{new}$, then let $\{u, v\} \in E'$ be an associated edge. Since R is an independent set it must be that $u \notin R$ or $v \notin R$. But this implies that in the graph $G - E^+(S)$, node i receives an edge from u or from v , so it has a positive in-degree. If, on the other hand, $i \in R$, then by our assumption that no node has degree of 0 in G' , we know that there is a $j \in V'$ such that $\{i, j\} \in E'$. Since R is an independent set, it holds that $j \notin R$. Thus, i receives an edge from j in the graph $G - E^+(S)$, hence it has a positive in-degree. This concludes the proof of the forward direction.

For the reverse direction, assume that there exists a subset $S \subseteq V$ such that $|S| = k$ and for every $i \in S$ it holds that $\deg_{G-E^+(S)}^-(i) > 0$. We will show that this implies the existence of an independent set of size r in G' . Let us denote by R the set of original nodes belonging in S , formally, $R := S \setminus V_{new}$. Since there are $m(n+1)$ new nodes in total, we can show that R has to contain at least r nodes. Indeed, $|R| = |S \setminus V_{new}| \geq |S| - m(n+1) = k - m(n+1) = r$.

We will now prove that R has to be an independent set in G' . For a contradiction assume that this is not the case, i.e., there exists an edge $\{u, v\} \in E'$ such that $u, v \in R$. Observe that there must be a new node $x_{\{u,v\}}^i$, for some $i \in [n+1]$, that is selected to S . Otherwise, if all the new nodes associated with the edge $\{u, v\}$ were not in S , the size of S would be too small, i.e., $|S| \leq |V| - (n+1) = m(n+1) + n - (n+1) = m(n+1) - 1 < k$. Since both u and v are in S , node $x_{\{u,v\}}^i$ has no incoming edges in $G - E^+(S)$, which contradicts our assumption proving that R is indeed an independent set. $\square$

A.2 Proof of Theorem 2

First, we observe that ABSORBRANK and ABSORBKATZ can be computed in polynomial time when applied to bipartite graphs. If the number of nodes with incoming edges, i.e., $|V_2|$, is at least k , then those rules simply coincide with their TOP counterparts. Otherwise, if $|V_2| < k$, then both ABSORBRANK and ABSORBKATZ output any subset of nodes of size k as the minimum centrality in such a subset is necessarily 1. Thus, in the remainder of the proof let us focus on functional graphs. Observe that on such graphs, ABSORBRANK coincides with ABSORBKATZ, simply because the two centralities are equivalent when the out-degree of every node is at most 1. Therefore, we focus on ABSORBRANK below.

It holds that if $G = (V, E)$ is a functional graph and $S \subseteq V$ is an arbitrary subset of nodes, then in the graph $G - E^+(S)$ each node $i \in S$ is a root of some in-tree, i.e., a connected functional graph without a cycle. This means that PageRank of i for $\alpha \rightarrow 1$ is just equal to the number of i 's predecessors in this graph. Hence, $\text{PR}_G^{\alpha \rightarrow 1}(S) = \min_{i \in S} |\text{Pred}_{G-E^+(S)}(i)|$. Therefore, we get that ABSORBRANK chooses a subset of nodes $S \subseteq V$ that maximizes the minimal number of predecessors of nodes from S in $G - E^+(S)$. To find such a subset, given parameters $p, \ell \in \mathbb{N}$ and a functional graph $G = (V, E)$, we ask whether there is a subset $S \subseteq V$ of size ℓ , such that each node $i \in S$ has at least p predecessors in $G - E^+(S)$. We will denote this decision problem by $\Pi(G, p, \ell)$. Deciding $\Pi(G, p, \ell)$ for all possible values of p (which are upper bounded by $|V|$) and selecting the set S that witnesses that $\Pi(G, p, k)$ is a *yes*-instance for the maximum possible value of p , solves the optimization version of ABSORBRANK.

It holds that $\Pi(G, p, \ell)$ can be solved in polynomial time for in-trees [32, 19]. We refer to such a procedure as $\text{alg-tree}(G, p, \ell)$. In what follows, we extend $\text{alg-tree}(G, p, \ell)$ first to arbitrary connected functional graphs (possibly with a cycle), and then to all functional graphs (possibly disconnected).

Suppose that G is an arbitrary connected functional graph. Observe that G can have at most one cycle and

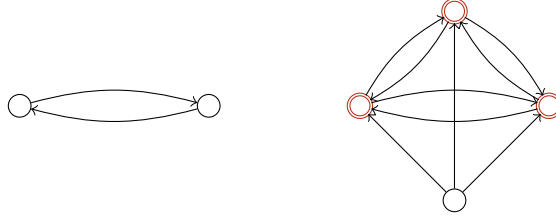


Figure 9: An example demonstrating that TOPRANK and TOPKATZ violate Clique-Entitlement (Proposition 3). Nodes with red double lines have the highest PageRank and Katz centrality.

we can assume that at least one node will be selected from the cycle (if no node is selected to S from the cycle, then we can take a selected node that is closest to the cycle, remove it from S , and instead select a node from the cycle; such a procedure will not decrease the number of predecessors in $G - E^+(S)$ of any node in S). Hence, for each node v in the cycle we can remove an outgoing edge of this node and check whether $\Pi(G - E^+(\{v\}), p, \ell)$ is a **yes**-instance using **alg-tree**. If this is the case for some node v in the cycle, we know that (G, p, ℓ) is a **yes**-instance as well, otherwise, it is a **no**-instance. Let us denote this procedure as **alg-connected**.

Finally, suppose that G is an arbitrary functional graph. By c we denote the number of its components and by $G_1, G_2, \dots, G_c$ the components themselves. To solve $\Pi(G, p, \ell)$ we use a dynamic programming approach. Let A be a $c \times \ell$ array. The values in the first row of A reflect for which numbers $j \in [\ell]$ we can select j nodes in G_1 in such a way that each has at least p predecessors after the removal of their outgoing edges, i.e., $A[1, j] = 1$, if $\Pi(G_1, p, j)$ is a **yes**-instance, and $A[1, j] = 0$, otherwise. Since G_1 is a connected functional graph, $\Pi(G_1, p, j)$ can be decided using **alg-connected**. Then, for each consecutive row $i \in \{2, 3, \dots, c\}$, the values reflect whether we can select such j nodes in $G_1, G_2, \dots, G_i$ jointly. Thus, $A[i, j] = 1$, if there exists a $j' \in [j]$ such that $A[i-1, j'] = 1$, and $\Pi(G_i, p, j-j')$ is a **yes**-instance, i.e., we can select j' nodes in $G_1, G_2, \dots, G_{i-1}$, and remaining $j-j'$ nodes in G_i (we assume that $\Pi(G', p, 0)$ returns **yes** for every graph G'). Finally, $\Pi(G, s, \ell)$ is a **yes**-instance, if and only if, $A[c, \ell] = 1$. $\square$

A.3 Proof of Proposition 3

In Figure 9 we present a graph that demonstrates that TOPRANK and TOPKATZ violate Clique-Entitlement, for $k = 3$. The graph consists of 2 components, one of size 2 that is strongly connected, say G_1 , and one of size 4 that is only weakly connected, say G_2 . Under the examined axiom, the nodes from G_1 are entitled to 1 representative in the selected committee. However, the three nodes of maximal PageRank and Katz centrality all belong to G_2 . As a result, no node from G_1 will be selected under TOPRANK or TOPKATZ. $\square$

A.4 Proof of Theorem 4

Let us start by showing that ABSORBKATZ violates Clique-Entitlement. To this end, we will use the Katz recursive formula that is an equivalent definition of Katz centrality [24]. It relates the centrality of a node to the centrality of its direct predecessors as follows:

$$K_G^\alpha(v) = \alpha \sum_{(u,v) \in E} K_G^\alpha(u) + 1$$

Let us denote an n -clique by K_n . Simple calculations based on the above formula show that

$$K_{K_n}^\alpha(v) = \frac{1}{1 - (n-1)\alpha}.$$

Hence, Katz centrality of a node in a clique is strictly increasing with the size of the clique for a given α . Moreover, observe that if we remove outgoing edges of a node in an n -clique, then Katz centrality

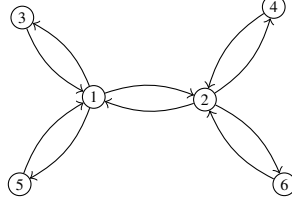


Figure 10: A component of the graph used in the proof that ABSORBRANK does not satisfy Component-Entitlement (Theorem 4).

of the remaining nodes will be equal to the centrality in $(n - 1)$ -clique—the existence of this node does not affect the number of walks ending in other nodes. Thus, if we have a graph in which one third of all nodes form clique K_n and two thirds form clique K_{2n} , then, for $k < n$, ABSORBKATZ would select only nodes from the larger clique. This clearly violates Clique-Entitlement.

Now, let us prove that ABSORBRANK violates Component-Entitlement. To this end, consider graph $G = (V, E)$ that consists of a clique of 24 nodes and disconnected from that a subgraph identical to that presented in Figure 10. Let us set $k = 15$, i.e., we want to select half of the nodes. By Component-Entitlement this would mean that we need to select 3 nodes from the structure in Figure 10. However, ABSORBRANK will select at most 2 from there. Specifically, if we select 2 from the structure and 13 from the clique, we get PageRank 2.98 for each in the structure and 1.83 for each from the clique, resulting in a minimum of 1.83. On the other hand, say that we select 3 from the structure and 12 from the clique. Then, we get PageRank of 1.90 for each in the clique. If we pick the 2 middle nodes from the structure or 1 middle and a neighbor of it, then we have a PageRank of 1. If we pick 1 middle and 2 non-neighboring (e.g. 1 and 4,6) then the minimum PageRank equals 1.33. If we pick 0 from the middle (e.g. picking 3,5,6) then the minimum PageRank equals 1.78. As a result, ABSORBRANK will pick at most 2 from the structure, since, adding any node as third, results in lower minimal PageRank than 1.83.

In the remainder of the proof, let us show that ABSORBRANK satisfies Clique-Entitlement. To this end, consider arbitrary graph $G = (V, E)$, in which there exists a component S such that $G[S]$ is a clique. Let $W \subseteq V$ be an arbitrary subset of nodes of size $|W| = k < n$. Also, let us denote the number of nodes from S in W by $\ell = |W \cap S|$. We will first prove the following bounds on PageRanks of nodes in W in the graph with outgoing edges of W removed:

$$\text{PR}_{G-E^+(W)}^{\alpha \rightarrow 1}(i) = |S|/\ell, \quad \text{for every } i \in W \cap S, \text{ and} \quad (4)$$

$$\text{PR}_{G-E^+(W)}^{\alpha \rightarrow 1}(j) \leq (n-|S|)/(k-\ell), \quad \text{for some } j \in W \setminus S. \quad (5)$$

Then, we will show that both formulas imply that whenever the Clique-Entitlement condition is not satisfied, i.e., $\ell < \lfloor k \cdot |S|/n \rfloor$, then the minimum PageRank of nodes from W in $G - E^+(W)$ can be increased. Since ABSORBRANK maximizes this minimum, this will mean that ABSORBRANK satisfies Clique-Entitlement.

We note that for every node $v \in V$ with a successor that does not have outgoing edges, Equation (1) is also well defined for $\alpha = 1$ (as the probability that the random walk returns to v after visiting it is strictly less than 1, the expected number of visits in the walk is finite). Thus, since we will consider only such nodes, to simplify our calculations, we will consider PageRank with $\alpha = 1$.

To prove Equation (4), we will use PageRank's recursive formula, which is an equivalent definition of PageRank [30]. The formula relates PageRank of a node with PageRanks of the nodes from which it receives incoming edges as follows

$$\text{PR}_G^\alpha(v) = 1 + \alpha \cdot \sum_{(u,v) \in E} \frac{\text{PR}_G^\alpha(u)}{\deg^+(u)}.$$

Consider an arbitrary unselected node in the clique, i.e., $v \in S \setminus W$. Since in graph $G - E^+(W)$, node v receives one edge from every other node in $S \setminus W$ and each such node has an out-degree of $|S| - 1$, by PageRank's recursive formula

$$\text{PR}_{G-E^+(W)}^1(v) = 1 + \frac{1}{|S| - 1} \cdot \sum_{u \in S \setminus W \setminus \{v\}} \text{PR}_{G-E^+(W)}^1(u).$$

Now, when we sum this equation side-wise for all $v \in S \setminus W$, on the right-hand side of the equation, $\text{PR}_{G-E^+(W)}^1(u)$ for each node u will appear $|S \setminus W| - 1 = |S| - \ell - 1$ times. Thus,

$$\sum_{v \in S \setminus W} \text{PR}_{G-E^+(W)}^1(v) = |S| - \ell + \frac{|S| - \ell - 1}{|S| - 1} \sum_{u \in S \setminus W} \text{PR}_{G-E^+(W)}^1(u).$$

Moving the sum to one side and dividing by $1 - (|S| - \ell - 1)/(|S| - 1) = \ell/(|S| - 1)$, we get

$$\sum_{v \in S \setminus W} \text{PR}_{G-E^+(W)}^\alpha(v) = (|S| - \ell) \frac{|S| - 1}{\ell}.$$

Then, take an arbitrary node $i \in S \cap W$ and observe that it receives one edge from every node in $S \setminus W$. Thus,

$$\text{PR}_{G-E^+(W)}^1(i) = 1 + \frac{1}{|S| - 1} \cdot \sum_{v \in S \setminus W} \text{PR}_{G-E^+(W)}^\alpha(v) = 1 + \frac{|S| - \ell}{\ell} = \frac{|S|}{\ell}.$$

Hence, Equation (4) indeed holds.

Next, let us prove Inequality (5). To this end, let us denote by T the set of all predecessors of nodes in $W \setminus S$ that are not selected to W themselves, i.e., $T = \{v \in V : \exists u \in W \setminus S, v \in \text{Pred}(u)\} \setminus W$. Then, let us sum PageRank's recursive formula sidewise, for all nodes $v \in T$. We obtain

$$\sum_{v \in T} \text{PR}_{G-E^+(W)}^1(v) = |T| + \sum_{(u,v) \in E: v \in T} \frac{\text{PR}_{G-E^+(W)}^1(u)}{\deg^+(u)}.$$

For every $u \in T$, let $d_w(u)$ be the number of edges that go to nodes in W (possibly $d_w(u) = 0$), while by $d_r(u)$ let us denote the number of its remaining outgoing edges, i.e., $d_r(u) + d_w(u) = \deg^+(u)$. Observe that every node that has an outgoing edge to a node in T must be in T itself. Thus, the set of edges between nodes in T is exactly the same as the set of all edges counted in $d_r(u)$ for some $u \in T$. This gives us

$$\sum_{v \in T} \text{PR}_{G-E^+(W)}^1(v) = |T| + \sum_{u \in T} \left(\frac{d_r(u)}{\deg^+(u)} \text{PR}_{G-E^+(W)}^1(u) \right).$$

Moving all PageRanks to the left-hand side, we get

$$\sum_{v \in T} \left(\frac{d_w(v)}{\deg^+(v)} \text{PR}_{G-E^+(W)}^1(v) \right) = |T|.$$

Now, let us sum the PageRank's recursive formula for all $j \in W \setminus S$. We get

$$\sum_{v \in (N \setminus S) \cap W} \text{PR}_{G-E^+(W)}^1(v) = k - \ell + \sum_{(u,v) \in E: v \in (N \setminus S) \cap W} \frac{\text{PR}_{G-E^+(W)}^1(u)}{\deg^+(u)}.$$

Observe that the set of all incoming edges to nodes in $W \setminus S$ is exactly the set of edges that are counted in $d_w(u)$ for some $u \in T$. Thus, we get

$$\sum_{v \in (N \setminus S) \cap W} \text{PR}_{G-E^+(W)}^1(v) = k - \ell + \sum_{u \in T} \left(\frac{d_w(v)}{\deg^+(v)} \text{PR}_{G-E^+(W)}^1(v) \right) = k - \ell + |T| \leq k - \ell + (n - |S| - (k - \ell)) = n - |S|.$$

Then, Inequality (5) follows from the pigeonhole principle.

Next, let us show that if $\ell \leq \lfloor k \cdot |S|/n \rfloor$, then a node in W with minimum PageRank is in set $V \setminus S$. To this end, observe that

$$\frac{|S|}{\ell} \geq \frac{|S|}{\lfloor k \cdot |S|/n \rfloor} \geq \frac{n}{k}.$$

Multiplying both the nominator and the denominator by $(1 - |S|/n)$, we obtain

$$\frac{|S|}{\ell} \geq \frac{n(1 - |S|/n)}{k(1 - |S|/n)} = \frac{n - |S|}{k - k \cdot |S|/n} \geq \frac{n - |S|}{k - \lfloor k \cdot |S|/n \rfloor} \geq \frac{n - |S|}{k - \ell}.$$

Thus, by Equation (4) and Inequality (5), indeed, if $\ell \leq \lfloor k \cdot |S|/n \rfloor$, then $\min_{i \in W} \text{PR}_{G-E^+(W)}^1(i) \leq \frac{n - |S|}{k - \ell}$. This means, that as long as the inequality is strict, i.e., $|S \cap W| < \lfloor k \cdot |S|/n \rfloor$, the minimum can be increased by removing a node with minimal PageRank from W and adding to W another node from the set S . This means that **ABSORBRANK** satisfies Clique-Entitlement. $\square$

A.5 Proof of Theorem 5

Our proof is based on the fact that the Method of Equal Shares satisfies a property called *priceability* [34]. The property portrays an election as a market in which the voters who control equal parts of a budget will jointly pay for candidates.

Given an election profile (V, C, μ) and requested committee size k , a price system is a pair $\mathcal{P} = (b, p)$, where $b \in \mathbb{R}$, $b \geq k$ is an *initial budget* (where each voter controls an equal part of the budget, i.e., b/n) and $p = (p_i)_{i \in V}$ is a sequence of *payment functions* where for each $i \in V$ the function $p_i : C \rightarrow \mathbb{R}_{\geq 0}$ represents how much money voter i spends for each candidate. We will assume that each candidate costs 1 unit of money. Then, a committee W is *supported by the price system* $\mathcal{P}$ if the following conditions hold:

1. The voters do not pay for candidates they do not support, i.e., $\mu_i(c) = 0$ implies that $p_i(c) = 0$ for every $i \in V$ and $c \in C$.
2. The sum of payments by i does not exceed their budget, i.e., $\sum_{c \in W} p_i(c) \leq b/n$.
3. Each elected candidate is fully paid, i.e., $\sum_{i \in V} p_i(c) = 1$ for every $c \in W$.
4. The voters do not pay for unelected candidates, i.e., $p_i(c) = 0$ for every $i \in V$ and $c \in C \setminus W$.
5. For each unselected candidate, its supporters do not have enough money to buy it, i.e., $\sum_{i \in V: \mu_i(c) > 0} (b/n - \sum_{c' \in W} p_i(c')) < 1$.

A committee W is *priceable* if there is a price system $\mathcal{P}$ that supports it. Peters et al. [34] observed that MES always outputs a priceable outcome. We will use this fact to show that it satisfies Subgraph-Entitlement.

Assume otherwise and take an arbitrary graph $G = (V, E)$, constant $k < n$, and utility function $\mu_G : V \times V \rightarrow \mathbb{R}_{\geq 0}$ such that $\mu_G(u, v) > 0 \Leftrightarrow u \in \text{Pred}(v)$ and there exists a subset S with $G[S]$ being strongly connected for which $|(S \cup \text{Succ}(S)) \cap W| \leq \lfloor k \cdot |S|/n \rfloor - 1$, where W is an outcome of MES on election (V, V, μ_G) with committee size k .

Since W is priceable, we know that there exists a price system $\mathcal{P} = (b, p)$ that supports it. First, let us show that each node $v \in S$ can pay only for nodes in $(S \cup \text{Succ}(S)) \cap W$, i.e., $p_v(u) = 0$ for every $v \in S$ and $u \notin (S \cup \text{Succ}(S)) \cap W$. Indeed, for $u \notin S \cup \text{Succ}(S)$, there is no walk from v to u , hence by our assumption on μ_G and condition 1 of a priceable system, we get that $p_v(u) = 0$. Moreover, $p_v(u) = 0$ for $u \notin W$ by condition 4.

Thus, we can show the following lower bound on the money that is unspent by nodes in S .

$$\begin{aligned}
\frac{|S| \cdot b}{n} - \sum_{v \in S} \sum_{c \in C} p_v(c) &= \frac{|S| \cdot b}{n} - \sum_{v \in S} \sum_{c \in (S \cup \text{Succ}(S)) \cap W} p_v(c) \\
&\geq \frac{|S| \cdot b}{n} - \sum_{v \in V} \sum_{c \in (S \cup \text{Succ}(S)) \cap W} p_v(c) \quad (\text{increase the set over which we sum}) \\
&= \frac{|S| \cdot b}{n} - |(S \cup \text{Succ}(S)) \cap W| \quad (\text{by condition 3}) \\
&\geq \frac{|S| \cdot b}{n} - \left(\left\lfloor k \cdot \frac{|S|}{n} \right\rfloor - 1 \right) \\
&\quad (\text{as } S \text{ is supposed to witness Subgraph-Entitlement violation}) \\
&\geq \frac{|S| \cdot b}{n} - k \cdot \frac{|S|}{n} + 1 \quad (\text{since } \lfloor x \rfloor \leq x) \\
&\geq 1. \quad (\text{since } b \geq k)
\end{aligned}$$

Now, let us show that there must exist a node in S that is unselected to W . Assume otherwise, i.e., $S \subseteq W$. Then, $(S \cup \text{Succ}(S)) \cap W = S \cup (\text{Succ}(S) \cap W)$. Moreover, since S is witnessing Subgraph-Entitlement violation, we get

$$|S| \leq |S \cup (\text{Succ}(S) \cap W)| = |(S \cup \text{Succ}(S)) \cap W| \leq \lfloor k \cdot |S|/n \rfloor - 1 \leq k \cdot |S|/n - 1.$$

However, this is a contradiction as this implies that $0 \leq (n-k)/n \cdot |S| \leq -1$.

Therefore, there is a node $i \in S \setminus W$. Since S is strongly connected, by assumption that $\mu_G(u, v) > 0 \Leftrightarrow u \in \text{Pred}(v)$, it must be that i is supported by all voters in S . On the other hand, as we have shown, these voters have together at least 1 unit of money, which contradicts condition 5 of a price system. $\square$